МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет имени
первого Президента России Б. Н. Ельцина»

# ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ КРИТЕРИЕВ И ТЕСТОВ. ПАРАМЕТРИЧЕСКИЕ МЕТОДЫ СТАТИСТИКИ

**Методические указания к выполнению**
**практического задания № 2**

Екатеринбург
2024

## Содержание

## Введение

На прошлой лабораторной работе мы изучили базовые средства работы с временными рядами, а также методы реализации собственных функций. Также там требовалось определить тип процесса, породившего данный ВР, что было достаточно трудным занятием, если не опираться на применение статистических критериев, которые будут рассматриваться в данной работе.

## 1. Задание на лабораторную работу

Результатом выполнения лабораторной работы является оформленный отчет в виде *Jupyter*-тетради, в котором должны быть представлены и отражены все нижеперечисленные пункты:

1) Сначала импортируйте в свой код нужные библиотеки, функции и т.д.

```
import numpy as np
import numpy.random as rand
import matplotlib.pyplot as plt
from scipy import signal
import scipy.stats as stats
%matplotlib inline
```

2) Создайте временной ряд, как частную выборку из нормального распределения:

```
x = rand.randn(10000)
t = np.linspace(3, 5, num = 10000)
plt.figure(figsize = (10, 5))
plt.plot(t, x)
plt.show()
```

3) Произведите оценку ВР на **стационарность**. Сначала используйте известный **KPSS-тест**. Для этого есть функция tsa.kpss(x,nlags="auto"), которая возвращает статистику теста **kpss_stat**, p-значение теста **p_value**, и другие полезные результаты (критические значения на разных перцентилях и т.д.).
4) Если KPSS-test в статистике близок к 0, то временной ряд является **стационарным** по КПСС-тесту, то есть **нулевая гипотеза** о тренд-стационарности ряда **принята**.
5) Если KPSS-test в статистике вернул значение существенно больше нуля, то временной ряд **не является стационарным** по КПСС-тесту, то есть **нулевая гипотеза** о тренд-стационарности ряда **отвергнута** и **принята альтернативная**.
6) Более точный показатель, на который следует обратить внимание – это **p-value**. Он, чаще всего, трактуется для данного КПСС-теста следующим образом: если значение **p-value** меньше **0.05** и стремится к нулю, то нулевая гипотеза **отклоняется**. Если значение больше **0.05**, то нулевая гипотеза **принимается**. При граничном значении **p-value** близком к **0.05** ответ является **неоднозначным** и гипотезу следует принимать или отвергать с большой осторожностью. На практике

значение **p-value** близкое к **0.05** обычно обозначает недостаточную длину исходной выборки анализируемых данных.

Для КПСС-теста граница **p-value** в **0.05** – неслучайна, так как он внутри программных библиотек ограничен сверху значением в **0.1**, из-за чего середина этого отрезка и определяет, когда принимать/отвергать гипотезу.

Но важно понимать – что для других статистических тестов *граница будет другой*. Поэтому более разумно всегда исходить из того, что чем ближе **p-value** к нулю – тем выше вероятность отвергнуть прямую гипотезу, и наоборот, чем ближе **p-value** в сторону единицы – тем выше вероятность принять прямую гипотезу. Но даже такое простое правило не всегда работает для статистических тестов – потому что бывают такие критерии, в которых отвергнуть гипотезу важнее ее принятия, из-за чего весь смысл значений меняется на противоположный. Это одна из больших проблем статистических тестов (что выдвигать на проверку, что принимать, что отвергать), которая упирается в саму программную реализацию теста. Поэтому, прежде чем использовать те или иные функции статистических библиотек, следует изучить их описание (*manual*, *help*), методы применения и получаемые результаты.

7) Скорее всего, случайная выборка нормального шума будет стационарной, так как не меняет со временем свои статистические характеристики, и КПСС-тест это покажет. Внесите явную нестационарность в этот ряд в виде тренда:

```
xv=x+(10*t**2-100*t+300)
plt.figure(figsize = (10, 5))
plt.plot(t, xv)
plt.show()
```

8) Примените к нему **KPSS-тест**. Сделайте выводы на основе анализа значений полученных **p-value**.
9) Возьмите теперь у исходного ВР и у нового нестационарного ВР две его половинки: пускай это будут **x1** & **x2** и **xv1** & **xv2**, например.

   Тогда проверьте с помощью **критерия Фишера** две половинки одного временного ряда на соответствие дисперсий, и затем проделайте то же самое для модифицированного временного ряда. Критерий Фишера можно реализовать через функцию **stats.f_oneway(x1, x2)**
10) Опишите получившиеся результаты с точки зрения «принятия»/«отвержения» гипотез и поясните получившиеся результаты.
11) Аналогично, теперь проверьте с помощью **критерия Стьюдента** две половинки исходного временного ряда на соответствие мат. ожиданий (при предположении о равных дисперсиях), и затем проделайте то же самое для модифицированного временного ряда. Используйте функцию **stats.ttest_ind(x1, x2)**
12) Опишите получившиеся результаты с точки зрения «принятия»/«отвержения» гипотез и поясните получившиеся результаты.
13) Самостоятельно найдите в библиотеке **Scipy.Stats** другие статистические критерии и тесты (в секции *Hypothesis Tests and related functions*). Примените **два** из них к Вашим выборкам ВР (или их половинкам), поясните – что они делают, какие гипотезы проверяют, какие результаты получились и почему.

14) Еще один очень большой недостаток статистических тестов состоит в том, что они относятся к **методам параметрической статистики**, и потому исходят из предположения, что работают с выборкой ВР, которая имеет какое-то отношение к **нормальному распределению**. Чем больше распределение отличается от нормального – тем более недостоверными будут являться результаты статистических тестов. Попробуем это продемонстрировать на простом примере: создайте две выборки ВР

**x = rand.randn(10000)**

**y = rand.rand (10000)**

Обратите внимание, что x – это выборка из **нормального распределения**, а вот y – это уже выборка из **равномерного распределения**. Примените к этим двум ВР статистический тест: **stats.ttest_ind(x, y)**

15) Скорее всего, значение **p-value** будет близко к нулю и гипотезу о равенстве мат. ожиданий следует отвергнуть. В самом деле – мат. ожидание **x** равно *0*, а вот мат. ожидание **y** равно *0.5*. Проверьте этот факт функцией **np.mean()**.

16) Но что будет, если у обеих выборок мат. ожидание сделать равным 0? Должны ли гипотеза быть принята, так как мат. ожидания равны? Должна ли гипотеза быть отвергнута, так как это выборки из разных распределений? Давайте проверим.

**xm = rand.randn(10000)**

**ym = rand.rand(10000)-0.5**

**stats.ttest_ind(xm, ym)**

17) Скопируйте 3 строчки кода выше и вставьте их в 5 разных ячеек, и запустите все их по порядку. При каждом вызове функций библиотеки ***rand*** будут выбираться разные выборки и получаться разные ВР, а значит и результат проверки во всех 5 ячейках будет отличаться. А будет ли результат отличаться существенно? Может они все будут близки к нулю? Или наоборот?

18) Сгенерируйте в этих 5 ячейках выборки так, чтобы получить не совпадающие значения выходных **p-value**. Поясните – в каких случаях гипотеза принималась или отвергалась, и попробуйте пояснить – почему.

19) Проделайте подобную работу (п. 16-18) для другого статистического теста **stats.f_oneway(xm, ym)** , а потом проделайте аналогичную работу для двух статистических тестов, которые Вы выбрали сами из п. 13.

20) Поясните полученные результаты.

## 2. Требования к оформлению отчета

Отчет в Jupyter-тетради должен обязательно содержать: номер лабораторной работы, ФИО студента, номер варианта (либо студенческий номер), номер группы, результаты выполнения работы с комментариями студента (комментарии пишутся после #) и изображениями.